

The Analysis of Spatial Networks by means of Link Communities

Carlo Drago

VERSIONE DEL 13/7/2016

Abstract

The main context in this work is the statistical analysis of the spatial networks. Our aim is to detect the spatial communities characterizing a network. The different communities can represent important social aggregations which are the place develops relevant social phenomena. In this context is particularly important to understand the network structure. Various community detection methodologies assume each node is belong a unique community. At contrary it could be important to specifically explore the spatial network structure and to check for the existence of some overlapping and nested data structures. The approach followed in this work, following Kalinka and Tomancak (2011) consider explicitly the link communities clusters because the possibility to detect nodes which present the characteristic to be an active member of multiple spatial communities. In this case, we are able to identify the complex structure of the network. In particular, we need to explore the overlapping and nested spatial community structure.

Carlo Drago, Università degli Studi Niccolò Cusano Telematica Roma, via Don Carlo Gnocchi 3, Roma, e-mail: carlo.drago@unicusano.it

1. L'identificazione di comunità su reti spaziali

Le reti sono un modo per rappresentare sistemi complessi posizionati nello spazio. In questo caso, all'interno della categoria delle reti spaziali il legame tra i singoli nodi può essere facilmente ricondotto a dei costi. Laddove quindi si realizzi questa assunzione di base una rete si può appunto definire spaziale (Barthelemy 2011).

Esempi di tali tipologie di rete possono essere i più diversi e possono riguardare reti di trasporto o anche di comunicazione. Internet, in particolare, risulta essere un rilevante esempio concreto di tali reti. In tutti questi casi sono importanti aspetti più prettamente fisici delle reti medesime come ad esempio la distanza tra i singoli nodi ma anche aspetti topologici relativi alla struttura della rete medesima.

La stessa struttura di rete risulta essere ubiqua nell'era dei big data e delle base dati di grandi dimensioni. In questo senso è importante notare come spesso la tipologia di rete che si riscontra assai frequentemente in natura presenta caratteristiche topologiche tipicamente non ovvie e per questa ragione considerate in letteratura "complesse" (Kim Wilhelm 2008). In questo senso definiamo una rete "complessa" quando presenta degli identificabili schemi a livello statistico e comunque diversa da reti casuali ("random graphs") i quali manchino di una qualunque struttura di questo tipo. Per una comparazione completa di queste reti si veda Van Der Hofstad (2009).

Le reti casuali o “random graph” rappresentano, quindi, reti a loro volta prive di una qualunque struttura come ad esempio un elevato coefficiente di clustering a livello topologico (in relazione agli indici di struttura delle reti si veda Wasserman e Faust 1994). In questo senso tali reti sarebbero definite come fortemente diverse da reti “small worlds” (Watts Strogatz 1998) le quali, a loro volta, sembrerebbero possedere delle caratteristiche specifiche che permettano ad esempio una più rapida diffusione di malattie ad esempio che all’interno di reti casuali.

E’ così possibile esplorare a livello di analisi dei dati le caratteristiche delle reti complesse identificando appunto le strutture dati più rilevanti e più significative anche per comprendere i fenomeni sociali sottostanti. Esistono, dunque, caratteristiche diverse che possono essere scoperte, ad esempio strutture gerarchiche della rete di riferimento o strutture fortemente modulari della stessa rete (Kim, Wilhelm 2008). Quest’ultima in questo senso si troverebbe ad essere formata da vari moduli distinti interconnessi tra loro. In generale l’ottimizzazione della modularità risulta essere assai rilevante in queste reti (Newman 2006).

In questo senso risulta essere particolarmente importante identificare la struttura delle interconnessioni di una rete proprio laddove queste siano particolarmente dense. All’interno di queste interconnessioni è infatti possibile identificare dei meccanismi sociali di relazione e di interscambio (Fortunato 2009). La struttura delle comunità di una rete risulta essere, quindi, importante perché misura la capacità di una rete di essere formata di tanti compartimenti che sembrerebbero da un lato comportarsi in maniera indipendente gli uni dagli altri (Krause et al. 2006) e dall’altro in maniera indipendente rispetto all’insieme della rete di riferimento.

Se quindi Barthelemy (2011) identifica, come detto, caratteristica delle reti spaziali che le interconnessioni tra i nodi possono essere considerati dei costi derivati, quindi risulta essere maggiormente importante identificare la struttura dei nodi che risulti essere identificabile come un gruppo omogeneo a partire dalle interconnessioni. Questa struttura della rete, dati anche i costi di riferimento per gli agenti economici, può avere un rilevante impatto su vari fenomeni sociali. Cruciale quindi l’analizzare statisticamente la rete alla ricerca della possibilità di identificare gruppi di nodi.

Questo è il caso dell’analisi delle comunità o identificazione di comunità o “community detection” (Fortunato 2009). In particolare l’analisi e l’identificazione delle comunità è utile a identificare degli schemi di nodi che presentino una determinata struttura di collegamento densa tra essi stessi come gruppo e meno densa considerando gli altri gruppi di nodi. In questo senso l’obiettivo risulta essere quello di identificare dei gruppi di nodi che presentino tale caratteristica dentro la rete portando all’identificazione della struttura della comunità dei nodi (o anche definita come “community structure” si veda in questo senso Girvan e Newman 2002).

Secondo Fortunato (2009) poi l’identificazione di comunità è importante in quanto permette di definire gruppi di nodi che presentino dei ruoli specifici nella rete di riferimento. Mediante l’analisi di tali ruoli è possibile anche ricostruire la struttura gerarchica della rete medesima. In questo senso l’identificazione di comunità permette di definire gruppi di nodi che possano essere considerati essi stessi come indipendenti rispetto alla rete di riferimento. Nello stesso modo tali gruppi di nodi potrebbero avere una qualche importante caratterizzazione a livello spaziale.

Varie metodologie sono state proposte in letteratura per l’analisi delle reti e l’identificazione delle comunità. In particolare sono stati proposti algoritmi di vario tipo per l’identificazione di nodi parte delle comunità di riferimento. Spesso, però, le differenze tra i singoli risultati dei diversi metodi ed algoritmi possono essere rilevanti (Leskovec Lang Mahoney 2011).

Le famiglie di metodologie sono state descritte in vari lavori come ad esempio in quello di Newman (2006). Un recente contributo in questo contesto che pone in rassegna vari lavori è quello di Wang et al. (2015) che presenta varie classi di approccio al problema dell’identificazione di comunità. Infine una specifica analisi comparativa a livello di performances è quella proposta da Lancichinetti e Fortunato 2009 che comparano diversi algoritmi di identificazione delle comunità. La suddivisione che propone Newman (2006) e la possibilità di ottenere risultati differenti porta alla necessità di considerare algoritmi che permettano di combinare informazione diversa. Drago (2016), infine, identificata la struttura delle comunità (“community structure”) propone di visualizzare ed esplorare la stessa mediante dati ad intervallo ottenibili a partire dalle singole comunità parte di una rete di grosse dimensioni.

In particolare negli ultimi anni sono state anche proposte metodologie che a partire da più metodi si arrivi ad un’unica definizione delle comunità identificate prevedendo una sorta di sintesi dei vari metodi o algoritmi utilizzati. Si parla in questo caso di approccio di consenso all’identificazione delle comunità della rete (“consensus community detection”). Vari approcci sono stati proposti in questo senso a partire da Lancichinetti e Fortunato (2012), Tepper Sapiro (2014) ed anche Fardad (2015). Si veda in questo contesto

Drago e Balzanella (2013) che, partendo da un numero di algoritmi diversi, usano una metodologia statistica per sintetizzare i risultati ottenuti dai diversi modelli identificando le comunità di riferimento. Come infatti è stato riscontrato è possibile ottenere risultati diversi facendo uso di algoritmi diversi, dovuti spesso a caratteristiche specifiche del singolo algoritmo (Leskovec Lang Mahoney 2011). In questo senso approcci come quelli proposti dagli autori puntano a ottenere una sintesi a partire dai risultati diversi.

In questo senso tutte queste tecniche sono particolarmente utili in quanto all'interno di reti geografiche permettono di identificare i gruppi di nodi che fanno parte delle diverse comunità.

Un altro contributo importante è quello di Palla et. al. (2005) che permette di identificare all'interno della rete anche nodi che fanno simultaneamente parte di più comunità. Gli autori riscontrano la presenza di significative sovrapposizioni nelle reti. In particolare laddove gli approcci classici nell'identificazione di comunità permetteva di identificare i gruppi di nodi a livello deterministico un tale approccio permette di identificare quei gruppi di nodi che presentano delle caratteristiche di comunità sovrapposte. L'analisi di tali meccanismi risulta quindi essere particolarmente importante in quanto è possibile riscontrare nelle reti geografiche questo fenomeno di sovrapposizione delle comunità per alcuni nodi di riferimento.

Una ulteriore metodologia, infine, è quella proposta da Ahn et al. (2010) Kalinka e Tomancak (2014) che permette non solo di visualizzare la struttura delle comunità in relazione ai nodi ma anche in relazione a quelle che sono le interconnessioni della rete medesima. In questo senso la procedura di identificazione delle comunità di nodi a livello statistico viene effettuata mediante la classificazione dei legami tra i diversi nodi considerati. Questa si può considerare una rilevante differenza con i metodi considerati come classici nell'identificazione delle comunità a livello di nodi. In particolare è riconosciuto che considerare le connessioni insieme ai nodi permette di migliorare il processo di identificazione delle comunità (Kalinka 2014 ed Evans Lambiotte 2009)

Ciò ha particolare importanza nelle reti spaziali proprio per il ruolo delle interconnessioni. In questo modo è poi possibile identificare i nodi che sono "chiave" dentro la rete ed anche quelle parti di rete che risultano presentare rilevanti caratteristiche di comunità sovrapposte. In questo senso le comunità sovrapposte.

Un'altra caratteristica delle reti geografiche identificabile con tale metodologia è la capacità per alcune parti di rete di presentare dei nodi della rete che a loro volta possono contenere delle altre comunità. L'utilizzo di un tale algoritmo può quindi essere molto utile a livello interpretativo per identificare tali strutture.

2. Studio di simulazione

Si considerano varie reti geografiche simulate ottenute mediante il pacchetto Igraph in R (Csardi Nepusz 2006). In particolare sono state considerate reti ottenute con l'algoritmo di Barabasi Albert (2002) ed anche di Leskovec et al. (2007). Le reti simulate sono state successivamente analizzate a livello quantitativo e per ciascuna di esse si sono considerati i vari metodi di identificazione delle comunità facendo uso di approcci di identificazione delle comunità di nodi o delle comunità di legami (come una differente metodologia). Per ciascun modello si sono seguiti vari approcci identificando quello che permettesse di meglio rappresentare la rete in comunità di riferimento.

In questo primo caso la rete è stata ottenuta mediante il modello Barabasi ed Albert (2002) che permette di rappresentare una rete di tipo complesso utile all'analisi ed alla identificazione delle diverse comunità di riferimento. Una tale rete viene definita spaziale proprio in quanto siamo in grado di associare alle singole interconnessioni tra i singoli nodi una distanza e quindi dei costi. La rete risulta essere basata su 50 nodi. La rete è visualizzata in fig. 1

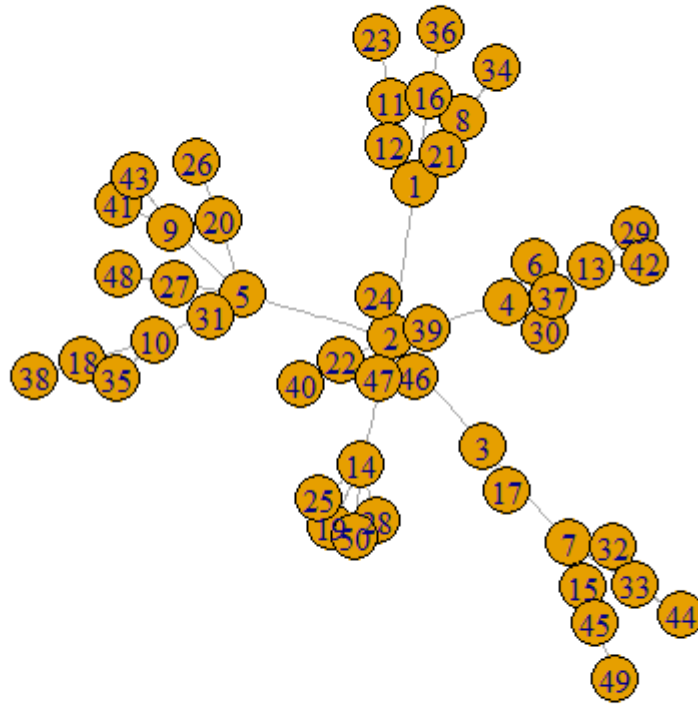


Fig. 1 Rete simulata 1: Modello Barabasi Albert

A partire dalla rete iniziale è possibile analizzare la struttura della rete andando a identificare i singoli nodi con degli indici di centralità. In questo senso si segue Wassermann e Faust (1994). Di tali indici di centralità si considerano primariamente la betweenness e la misura di Freeman. Infine si passa all'utilizzo di vari algoritmi di identificazione delle comunità tra cui si considerano vari approcci. In particolare nella fig. 2 si visualizza il risultato ottenuto dall'algoritmo Walktrap (Pons Latapy 2005).

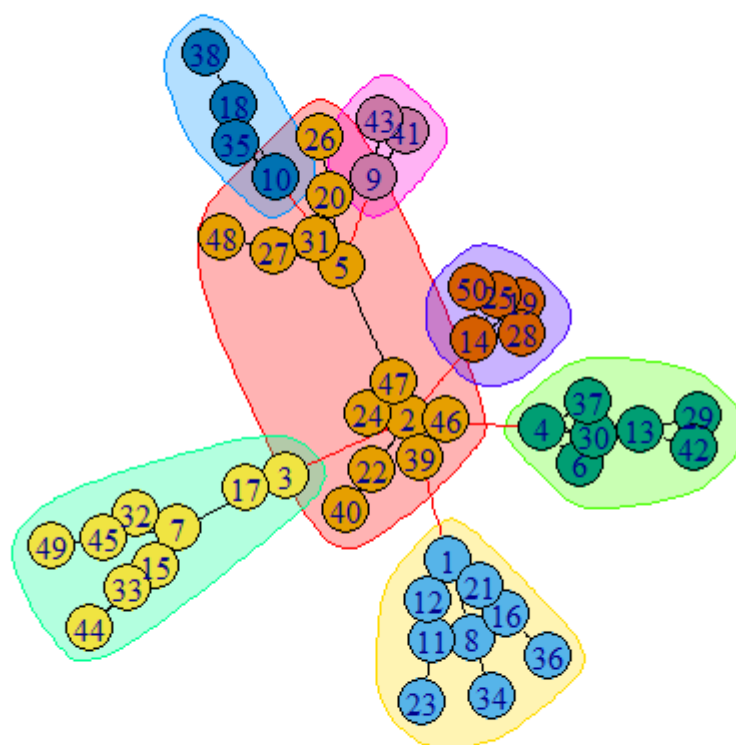


Fig. 1 b Rete simulata 1: comunità di nodi identificate (nostre elaborazioni su reti simulate)

La rappresentazione delle comunità sembra soddisfacente in quanto permette di identificare gruppi di nodi i quali sembra realistico ipotizzare anche una interazione spaziale tra di loro.

Considerando una rete diversa ottenuta con un algoritmo differente simulazione “Forest Fire” (Leskovec et al. 2007) il medesimo algoritmo ha reso necessario il considerare l’identificazione della struttura di comunità mediante l’algoritmo di Ahn et al. (2010) implementato in R da Kalinka e Tomancak (2011). A questo punto è stato anche possibile altresì identificare i nodi fondamentali parte delle diverse comunità della rete.

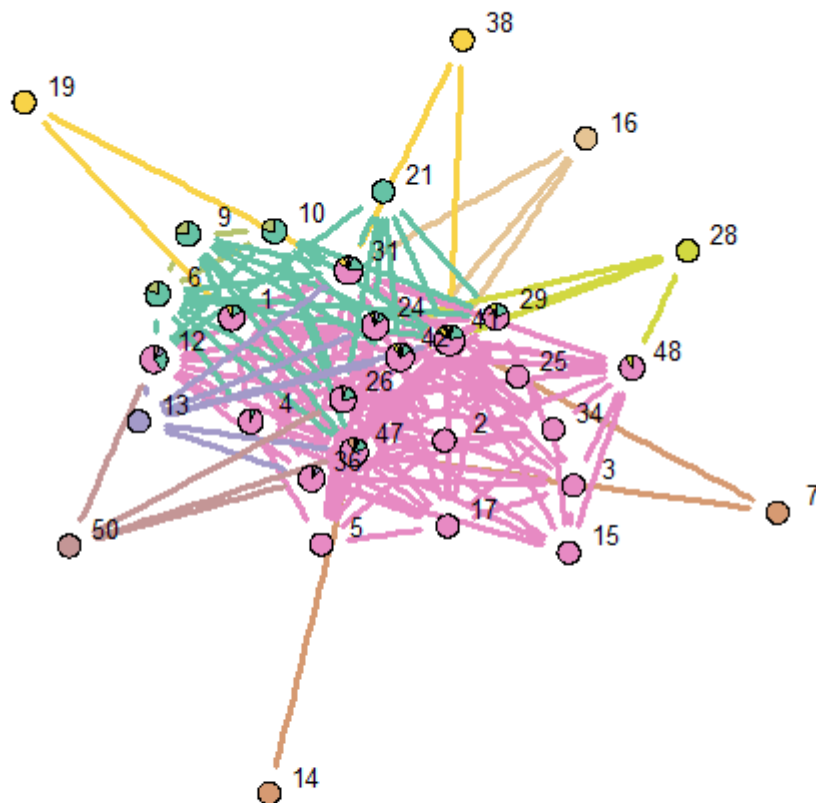


Fig. 2 Rete simulata 2: Modello Forest Fire e comunità di legami identificate (nostre elaborazioni su reti simulate)

3. Conclusioni

L'analisi e l'identificazione delle comunità risulta essere particolarmente importante nell'analisi delle reti spaziali. In particolare è possibile riscontrare spesso la necessità di identificare gruppi di nodi che presentano delle interconnessioni tra di essi maggiormente dense. In questo senso vari algoritmi sono stati proposti per l'identificazione di tali comunità di nodi, sia nell'identificazione di gruppi maggiormente densi al loro interno e poco densi all'esterno che nella scoperta di gruppi che presentino una struttura gerarchica con un gruppo di nodi contenuto all'interno di altri. La letteratura in ambito dell'analisi di rete ha negli ultimi anni proposto approcci che tenessero conto altresì della presenza di reti sovrapposte e che a partire dalle sovrapposizioni identificassero quei nodi chiave a livello di sistema. L'algoritmo proposto da Ahn et al. (2010) nella versione proposta da Kalinka (2014) è stato comparato con altre metodologie per comprendere le sue performances a livello di reti spaziali.

Riferimenti bibliografici

Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307), 761-764.

- Albert R. Barabási, A. L. (2002). "Statistical mechanics of complex networks". *Reviews of Modern Physics* 74 (1): 47–97. doi:10.1103/RevModPhys.74.47. ISSN 0034-6861.
- Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1), 1-101.
- Csardi G, Nepusz T (2006): The igraph software package for complex network research, *InterJournal, Complex Systems* 1695. <http://igraph.org>
- Drago, C., & Balzanella, A. (2013, October). Consensus community detection: a nonmetric MDS approach. In *SIS Cladag 2013 9th Scientific Meeting of the Classification and the Data Analysis Group of the Italian Statistical Society*.
- Drago C. (2016) *Exploring Community Structure*
- Evans, T. S., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1), 016105.
- Fardad, M. (2015, December). On consensus-based community detection. In *2015 54th IEEE Conference on Decision and Control (CDC)* (pp. 1577-1582). IEEE.
- Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3), 75-174.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.
- Kalinka A. T. (2014) The generation, visualization, and analysis of link communities in arbitrary networks with the R package linkcomm. <https://cran.r-project.org/web/packages/linkcomm/vignettes/linkcomm.pdf>
- Kalinka, A. T., & Tomancak, P. (2011). linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*, 27(14).
- Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., & Taylor, W. W. (2003). Compartments revealed in food-web structure. *Nature*, 426(6964), 282-285.
- Kim, J., & Wilhelm, T. (2008). What is a complex graph?. *Physica A: Statistical Mechanics and its Applications*, 387(11), 2637-2652.
- Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5), 056117.
- Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific reports*, 2.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2.
- Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web* (pp. 631-640). ACM.

- Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23), 8577-8582.
- Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814-818.
- Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences* (pp. 284-293). Springer Berlin Heidelberg.
- Tepper, M., & Sapiro, G. (2014). All for one, one for all: Consensus community detection in networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1075-1079). IEEE.
- Van Der Hofstad, R. (2009). Random graphs and complex networks. Available on <http://www.win.tue.nl/rhofstad/NotesRGCN.pdf>, 11.
- Wang, C., Tang, W., Sun, B., Fang, J., & Wang, Y. (2015). Review on community detection algorithms in social networks. In *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)* (pp. 551-555). IEEE.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge university press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684), 440-442.